A Model for identification of Length of Longest Common Subsequence by SRLCS

Sumathy Eswaran
Research Scholar, Vels University, Chennai,
& Assistant Professor , Department of Computer Science and Engineering,
Dr MGR Educational and Research Institute University, Chennai, India

Dr.S P Rajagopalan
Professor Emeritus
Dr MGR Educational Research Institute University
Chennai, India

ABSTRACT

Longest Common Sequence problem is the most fundamental task in Computational Biology. This is not only a classical problem but also a challenging problem in bio sequences application. Many algorithms are being developed and these are discussed in terms of resource utilization efficiency. This paper proposes a model based on SRLCS algorithm [11] to obtain the possible length of Longest Common Sequence (LCS). The model accounts the length of the sequences under consideration, the identity and similarity between them. The model is obtained by regressing the LCS results on the training data set by SRLCS. The model so obtained is a simple linear expression which gives the predicted length of LCS. The possible Length of LCS between the given sequences is a sufficient heuristic for biologists in decision making. Often such a result is useful while working on homology finding.

General Terms

Computational Biology, Biosequences analysis

Keywords

Pair wise, Longest Common Subsequence length, Longest Common Subsequence model, Parallel Algorithm, Fast LCS, SRLCS.

1. INTRODUCTION

Bio sequences could be representing DNA, RNA, protein, Gene, Genome etc of an organism. Biologists are often interested to know the evolutionary, functional and structural relationship between organisms. Ab-initio methods of computational biology not only help reduce the time and resource required for lengthy laboratory process and also provide good direction for timely quality research by biologists.

Pair wise sequence alignment has long and fruitful history in computational biology. Sequence Alignment is the procedure of comparing two (pair wise) or more DNA or Protein sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. Multiple sequences alignment (MSA) is useful while comparing a sequence with a family of database of sequences. However pair wise alignment would suffice in one to one investigations and provide good predictions about the biological similarity for related sequence.

Longest Common Subsequence (LCS) problem determines the longest ordered subsequences between the given sequences. Thus sequence alignment can be approached through LCS identification. The LCS between two sequences is function of the length of query sequence (X), length of Reference sequence (Y), the identity Score(I) between the two sequences and the similarity Score (S) between the two sequences. i.e.

$$\left|LCS\right| = f_1(X) + f_2(Y) + f_3(I) + f_4(S) \text{ ....(1)}$$

Similarity score is the sum of the number of identical matches and conservative substitutions in a sequence alignment divided by the total number of aligned sequence characters. The identity score is the number of characters that match position wise in both sequences divided by the length of query sequence. Between any two sequences, the identity percentage reflects the existence and the extent of similarity too. However the converse relation does not hold true. Further, Identity score characterizes the quality of an alignment and the likelihood that it reflects homology.

2. RELATED WORKS

LCS problem is computationally complex when the sequences are longer. Classical method for finding LCS is Dynamic programming algorithms provided by Smith –Waterman [9] for Local alignment and Needleman-Wunsch [6] for global alignment. Dynamic programming solution complexity is O(nm) for both time and space for m sequences of length n. Decision tree model by Aho and et al.[1] gave lower bound of O(mn). Hirschberg[4] solution reduces the space complexity to O(m+n).

MLCS problem is NP-Hard. The time complexity of most algorithms for MLCS depends on the number of sequences. Lot of work has been done and many algorithms have been developed towards reducing the complexity. The parallel algorithms like FASTLCS[13,14] , EFPLCS[10] and parMLCS[7] gave near linear speed up for large number of sequences. FASTLCS complexity is O(|LCS(X,Y)|) for time complexity and max{4*(n+1)+4*(m+1), L} for space complexity. EFPLCS is 70% more efficient than FASTLCS in resource utilization of both memory and CPU. But EFPLCS complexity remains the same as FASTLCS.

Later many heuristic algorithms like THSB (Time Horizon Specialized Branching heuristic)[12] , Ant Colony Optimization [8], Beam search algorithms[2], SRLCS[11] have been developed. Heuristic algorithms play crucial role to identify LCS within reasonable time on large size sequences and the heuristic parameters used determine the solution quality. Solution quality can be set to an acceptable limit with reference to the problem in hand. As already said, LCS identification is the first step which helps design the experiments further required towards the goal.

SRLCS [11] algorithm by the same author is an MLCS parallel algorithm. When implemented using Parallel computing the complexity is O(|LCS(X1, X2,… Xn)|). i.e Its complexity is “Independent of the number of sequences n”. In the parallel implementation, each of the computing nodes brings out the MLCS with respect to the Initial Identical Pair (IIDP) assigned to it. Then the LCS(s) of maximum length is chosen and produced as LCS by the master computational node. Thus SRLCS [11] truly brings out the MLCS without any compromise or approximation while computationally efficient for longer sequences.

The Computational Biology requirements are growing so fast that sequential algorithms do not meet the expectations of biologists even with the so called powerful current generation computing resources. Even a clue to the solution by heuristic approaches helps the biologists to plan about next process towards their goal. So this paper proposes a model based on SRLCS [11] heuristic algorithm to know the probable LCS length between a pair of sequences. The model is derived and validated with protein sequences from pfamseq database [5].

3. EXPERIMENT

Sequences from pfamseq database were used. Protein Sequences from families PF03678.7, PF10786, PF10108.2, PF09805.2, PF9850, and PF10277.2 of about length 200 were taken as data set. Within each family, one sequence was chosen as query sequence and was compared on pair wise basis with others. In all 70 datasets were used. SRLCS identified the LCS for these 70 data sets. The experiment was carried out on a PENTIUM desktop system with 1GB main memory. The identity and similarity percentage scores of the datasets were collected using SSEARCH35 [9]. SSEARCH35 is available as part of FASTA [15]

3.1 CORRELATION WITH IDENTITY SCORE

Nucleotide Sequences of length in the range from 170 to 235 having identity percentage 28 to 100 were taken as data set. Using SRLCS algorithm the LCS identified is pictorially plotted in Fig. 1. It is observed that correlation exists between Identitiy and LCS Length.

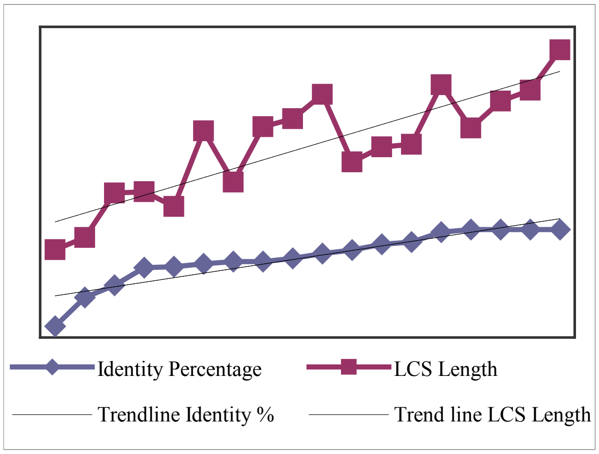

Figure 1. Identity Vs Length of LCS

The Correlation between identity score(x) and length of LCS (y) is calculated using the formula

$$CorrelationCoefficient\gamma = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

………………..(2)

The correlation Coefficient (Identity Score, Length of LCS) derived is 0.7576 implying strong positive correlation.

In fig .1. the trendlines have different slopes because of the fact that LCS is a function of 3 other factors as represented in equation (1).

3.2 . SRLCS MODEL

Similarity by definition has relevance to LCS. Positive correlation with LCS could be established as above.. This prompted to identify a model to calculate the length of LCS between given two sequences by a simple procedure. Therefore 19 datasets consisting of (|Xseq |, |Yseq|, Identity score(I), Similarity Score(S), Length of LCS) were taken as training data set. The first four items are inputs, based on which the LCS and its length are identified by SRLCS algorithm. On the training dataset, Regression Data analysis (|X seq |, |Y seq |, Identity score(I), Similarity Score(S), Length of LCS) is done to bring out the regression coefficients corresponding to each input variable in identifying the Length of LCS(output). Fig.2. thru Fig.5 , show the scatter charts for the LCS identified and the predicted LCS by the model with reference to the input variables used in regression. The regression model so derived for LCS length identification fits in as

$$|LCS| = 0.8Y - 0.017X + 2.89I - 2.43S$$ …….(3)

Where X is the length of Query sequence

Y is the length of Reference Sequence

I is identity percentage between X and Y sequences

S is the Similarity between the X and Y sequences

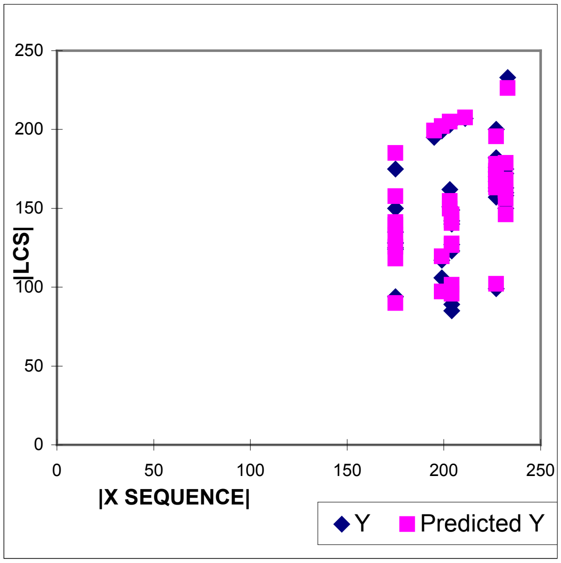

Figure 2.

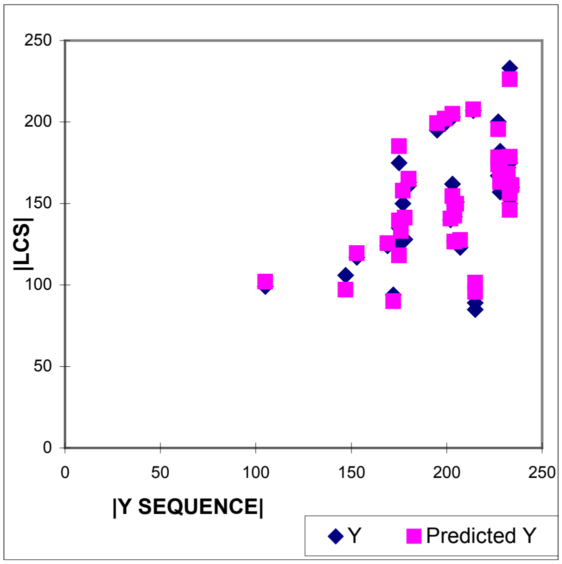

Figure 3

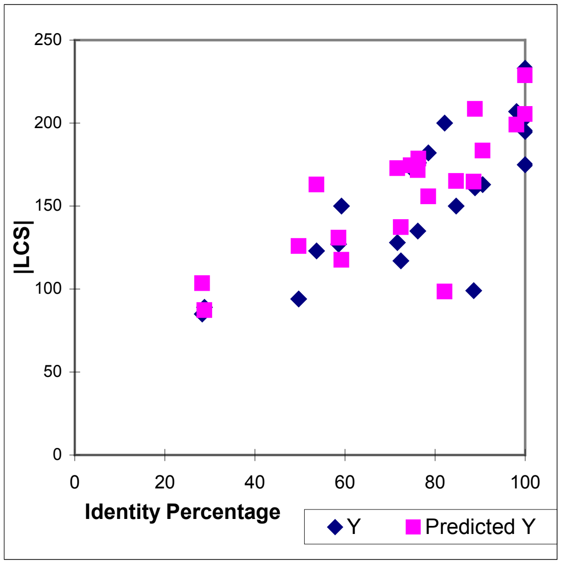

Figure 4

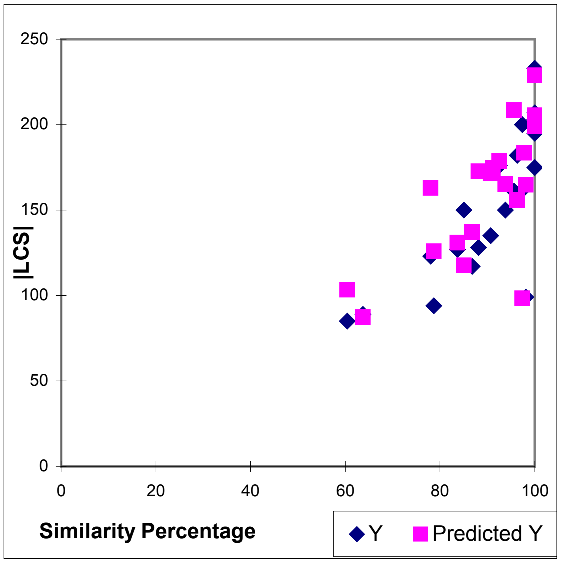

Figure 5

Scatter Chart for Predicted Vs.Actual LCS With reference to |X Sequence| in Figure 2 , |Y Sequence| in Figure 3, Identity Percentage in Figure 4, Similarity Percentage in Figure 5. The Legends Y and predicted Y are the LCS and predicted LCS.

3.3 VERIFICATION OF THE MODEL

22 other datasets used as Test data set-A. Using this regression model in equation (2) LCS length for test data set identified. Since the sample data size was 20, t-test was chosen for

verification. The hypotheses assumed are Null hypothesis (H_0) and Alternate Hypothesis (H_A).

Null hypothesis (H_0) = There is no difference in means between the Training sample (µ1) and test data set-A (µ2) sets. i.e . Ho= µ1- µ2 = 0.

Alternate Hypothesis (H_A) = There is difference in means between sample and test data sets.

t-test (Training sample, Test dataset-A, 2 tails, heteroscedastic) = 0.048597

The t-test table value corresponding to degree of freedom 41 with probability level of significance 0.05 for 2 tailed function is 2.021. The t-test value obtained is 0.4857 which is much below the table value. This proves the null hypothesis (H_0) to be true and alternate hypothesis to be false. The data sets are referenced in Table.1.

Table 1. Verification of SRLCS Model with Test Data Set - A

Training Data Result

Length of X seq	Length of Y seq	Identity	similarity	Regression o/p By SRLCS Model
204	215	28.8	63.7	97
204	207	53.7	78.0	128
232	233	59.2	85.0	147
232	233	63.5	85.8	157
232	234	65.4	86.3	163
199	147	65.8	86.6	94
232	233	66.1	86.7	163
204	202	67.0	87.2	140
227	228	67.1	85.5	165
204	204	67.5	87.7	142
175	175	68.0	90.1	115
232	232	69.8	87.9	170
204	204	70.0	89.2	145
175	169	70.8	88.1	123
203	205	72.2	90.7	149
175	178	72.3	85.9	140
227	228	72.8	89.9	170
232	233	73.4	88.0	181
227	227	73.6	88.5	175
203	203	78.2	95.5	153
227	227	87.7	96.5	197
199	199	100.0	100.0	202

Test dataset-A result

Length of X seq	Length of Y seq	Identity	similarity	Regression o/p By SRLCS Model
204	215	28.8	63.7	97
204	207	53.7	78.0	128
232	233	59.2	85.0	147
232	233	63.5	85.8	157
232	234	65.4	86.3	163
199	147	65.8	86.6	94
232	233	66.1	86.7	163
204	202	67.0	87.2	140
227	228	67.1	85.5	165
204	204	67.5	87.7	142
175	175	68.0	90.1	115
232	232	69.8	87.9	170
204	204	70.0	89.2	145
175	169	70.8	88.1	123
203	205	72.2	90.7	149
175	178	72.3	85.9	140
227	228	72.8	89.9	170
232	233	73.4	88.0	181
227	227	73.6	88.5	175
203	203	78.2	95.5	153
227	227	87.7	96.5	197
199	199	100.0	100.0	202

The same null and alternate hypotheses were tested with larger Test Data Set-B of size 51. This t-test table value for degree of freedom =70 and probability level of significance = 0.05 is 2.000. The **t-test (**Training sample, Test dataset-B, 2 tails, heteroscedastic) = 0.191568. Since this value is much below the t-table value the null hypothesis is proved to be correct. i.e. the model identified by SRLCS as in equation (3) is a fitting model for LCS length identification.

3.4 COMPARISON WITH OTHER TOOLS

The other Sequence alignment tools like Smith-Waterman SSEARCH35 [9] and MUSCLE [3] were run on the same dataset as SRLCS. The correlation between derived LCS of these methods found to be

The Correlation Coefficient (|SSEARCH35|, |SRLCS|) = 0.9921

The Correlation Coefficient (|MUSCLE|, |SRLCS|) = 0.9912

This implies that SRLCS identifies LCS as good as any other such methods in practice.

4. CONCLUSION

The SRLCS model for identifying the length of LCS simplifies the job of the computational biologist. The result obtained from the SRLCS model can be used as a heuristic by the biologists in his goal to find solution with the sequences under investigation.
Although the tests are done on sequential heuristic implementation, SRLCS is a parallel heuristic Algorithm for MSA. The complexity of SRLCS [11] is O(|LCS(X1, X2,… Xn)|) and is independent of the number of sequences. As SRLCS method is proved to have strong correlation to other methods in identifying LCS, it is believed that this model would be useful to biologists.

5. RERENCES

[1] A.Aho, D.Hirschberg and Jullman, 1976, Bounds on the Complexity of the Longest Common Subsequence Problem, J.Assoc.Comput.Mach., Vol. 23, No.1,1976

[2] Blum, C.; Blesa, M. J.; and L´opez-Ib´a´nez, M. 2009. Beam search for the longest common subsequence problem. *Comput. Oper. Res.* 36(12):3178–3186.

[3] Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, Vol. 32, No. 5, , 2004, 1792-1797

[4] Hirschberg, D. S. 1977. Algorithms for the longest common subsequence problem. *J. ACM* 24(4):664–675.

[5] Pfam 25.0 (March 2011, 12273 families), available at http://pfam.sanger.ac.uk/

[6] Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 48(3):443-453, 1970.

[7] Qingguo Wang, Dmitry Korkin and Yi Shang , Efficient Dominant Point Algorithms for the Multiple Longeset Common Subsequence(MLCS) problem , *IJCAI 2009.* 1494–1500.

[8] Shyu, S. J., and Tsai, C.-Y. 2009. Finding the longest common subsequence for multiple biological sequences by ant colony optimization. *Comput. Oper. Res.* 36(1):73–91.

[9] Smith t.F., Waterman M.S, Identification of common molecular subsequence, Journal of Molecular Biology, Vol.215, 1990

[10] Sumathy Eswaran , S.P.Rajagopalan, An Efficient Fast Pruned Algorithm for finding Longest Common Sequences in Bio Sequences, Annals.Computer Science Series, 8th Tome, 1st Fasc, page 137 – 150, 2010.

[11] Sumathy Eswaran, S.P.Rajagopalan , "Heuristic SRLCS Algorithm to determine the proper alignment strategy for Biosequences , International Journal of Research and Reviews in Information Technology (IJRRIT) ,Vol. 1, No. 2, ,1-7, June 2011, ISSN: 2046-6501

[12] Todd Easton, Abhilash Singireddy, 2008, A large neighborhood search heuristic for the longest common subsequence problem , *J Heuristics*(2008)14:271-283.

[13] Wei Liu, Lin Chen, A Fast Longest Common Subsequence Algorithm for Biosequences Alignment, IFIP vol 258, 2008.

[14] Yixi Chen, Andrew Wan and Wei Liu , A fast Parallel Algorithm for finding the Longest Common Subsequence of multiple biosequences , *BMC Bioinformatics 2006*, 7 (suppl 4): 54, ©2006 Chen et al; licensee BioMed Central Ltd.